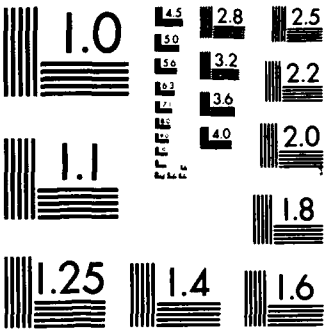
AD-A184 470 AN URN MODEL FOR THE MULTI-SAMPLE CAPTURE/RECAPTURE SEQUENTIAL TAGGING PR (U) CALIFORNIA UNIV BERKELEY OPERATIONS RESEARCH CENTER J G LEITE ET AL MAR 87 1/1

UNCLASSIFIED ORC-87-10 ARO-22735 4-MA DAAG29-85-K-0208 F/G 12/3 NL



AN URN MODEL FOR THE MULTI-SAMPLE CAPTURE/RECAPTURE SEQUENTIAL TAGGING PROCESS*

by

Jose Galvao Leite**
Carlos Alberto de Braganca Pereira***

ORC 87-10 March 1987

* Sponsored in part by CNPq, Brasilia, Brazil under contract #20771/85-MA, and the U.S. Army Research Office under contract #DAAG-85-K-0208. Reproduction in whole or in part is permitted for any purpose of the U. S. Government.

** IME - Universidade de Sao Paulo, CP20570, CEP01498, SP, Brazil.

*** IME - Universidade de Sao Paulo, CP20570, CEP01498, SP, Brazil. Visiting Research Engineer, Operations Research Center, University of California, Berkeley, CA 94720.

MASTER COPY - FOR REPRODUCTION PURPOSES

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered) AD-A184 470

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: ARO 22735.4-MA	2. GOVT ACCESSION NO.: N/A	3. RECIPIENT'S CATALOG NUMBER: N/A
4. TITLE (and Subtitle): An Urn Model for the Multi-Sample Capture/Recapture Sequential Tagging Process		5. TYPE OF REPORT & PERIOD COVERED: Technical
		6. PERFORMING ORG. REPORT NUMBER: ORC-87-10
7. AUTHOR(s): Jose Galvao Leite; Carlos Alberto de Braganca Pereira		8. CONTRACT OR GRANT NUMBER(s): DAAG-85-K-0208
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Operations Research Cemter, 3115 Etcheverry Hall, University of California, Berkeley, CA 94720		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: N/A
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, Post Office Box 12211, Research Triangle Park, NC 27709		12. REPORT DATE: March 1987
		13. NUMBER OF PAGES: 10
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report): Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

CMRR sampling process; capture/recapture sequential sampling process; random allocation; alocation process; sufficient statistic.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(SEE REPORT)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AN URN MODEL FOR THE MULTI-SAMPLE CAPTURE/RECAPTURE SEQUENTIAL TAGGING PROCESS

José Galvão Leite and Carlos Alberto de Bragança Pereira

IME-Universidade de São Paulo, CP 20570, CEP 01498, SP, Brazil

Key words and Phrases:CMRR sampling process; capture/ recapture sequential sampling process; random allocation; allocation process; sufficient statistic.

ABSTRACT

The probability distribution associated with the multisample CMRR generalized sequential sampling process are obtained by using an analogy with a single urn model. Some statistical features are also discussed.

1. INTRODUCTION

The Capture/Marc/Release/Recapture (CMRR) sampling process is used whenever informative data must be obtained in order to estimate the unknown size, N, of a finite (and closed) population. The sampling design for such process is described ~~here~~. in this document.

Consider a population of finite size, N, such that during the study time it changes neither in size nor in form; that is, the population is **closed** during the study time. From this population, k (k is fixed and ≥2) random samples (without replacement) are sequentially selected in the following manner:

The first random sample of (fixed) size m_1 (≥1) is drawn, without replacement. After the sample units are **marked** and the number $m_1=U_1$ is **recorded** they are **returned** to the population before the second sample is drawn. Next, for each j (≥2), the j^{th} random sample of (fixed) size m_j (≥1) is drawn, without replacement. The sample units marked in earlier selected samples are immediately returned to the population. The remaining U_j unmarked sample units are returned after being

marked. The numbers m_j and U_j are recorded. After the k samples have been obtained, the data

$$D_k = (U_1,...,U_k)$$

is observed. Note that the number of distinct population units selected in the whole sample process is

$$T_k = U_1+...+U_k \ .$$

The objective of the present paper is to obtain the probability laws of D_k and T_k by using an equivalent urn model. By urn model we mean random allocations of balls to urns.

The CMRR sampling scheme has a long reference list (see Seber, 1986) which starts with Craig (1953) and Goodman (1953), although, a related problem was described earlier by Good (1950, p.73). The majority of the papers [viz.Samuel (1968) and Sen (1982), among others] consider only the one-by-one case (i.e., $m_1=...=m_k=1$) and none of them presents the probability law of D_k, the raw data. We believe that these restrictions are in fact necessary when difference equations (the tool of many authors) are to be used to obtain these laws. The distribution of T_k, for the general case of m_j different from 1 for some j, is described in Johnson & Kotz (1977, Section 5.3) where an analogy with the committee problem is used. Also, in this text, no reference to D_k is made. In fact, for inferences about N, it is enough to consider only T_k since it is a sufficient statistic for N in relation to D_k, as show in Section 3. Note also that T_k and N are both positive integer numbers while D_k is a non-negative integer vector of order k. We end this section noticing that the sequence $(U_i)_{i>1}$ is not an exchangeable sequence which implies that it is not a sequence of conditionally independent and identically distributed random variables. Hence, T_k is sufficient in the broad sense. That is, the conditional distribution of D_k given T_k is the same for every possible N.

2. ANALOGY AND NOTATION

Consider an imaginary one-to-one correspondence between population units and urns; that is, a different urn is assigned to each one of the N population units. Also consider $m=m_1+...+m_k$ balls numbered in the following way: m_1 with the number one, m_2 with the number two, and so on up to m_k with the number k.

To select, without replacement, m_1 population units to be marked corresponds to randomly allocating to the urns the m_1 one-numbered balls, in such a way that no urn receives more than one of these balls. To select, without replacement, the second sample of m_2 population units corresponds to randomly allocating to the urns the m_2 two-numbered balls, in such a way that no urn receives more than one of these balls. To count the number U_2 of unmarked sample units (to be marked) is equivalent to counting the urns, among the m_2 ones that received the two-numbered balls, with only one ball. Sequentially following this analogy, consider the j^{th} sample ($j>1$). To select, without replacement, the j^{th} sample of m_j population units corresponds to randomly allocating to the urns the m_j j-numbered balls, in such a way that no urn receives more than one of these balls. To count the number U_j of unmarked sample units (to be marked) is equivalent to counting, among the m_j urns that receive the j-numbered balls, the ones with only one ball. (Note that at the end of this allocation process, it may happen that many urns are empty, some have only one ball, and so on up to a very few having k balls.)

Following the above analogy, in the remaining part of the present paper, the vector $D_k=(U_1,...,U_k)$ represents indifferently either the data obtained by the CMRR scheme described in Section 1 or the data obtained by the urn scheme described above. Before presenting the probabilities of interest, we introduce the notation used.

As usual the indicator function of a set A is represented by $I_A(x)$. Also, let $\mathbf{N}^*= \{0,1,...\}$ be the set of non-negative integers.

In general, for $j\geq 1$, the random vector $D_j=(U_1,...,U_j)$ has its observed vector represented by $d_j=(u_1,...,u_j)$. Analogously, for $T_j=U_1+...+U_j$, we have $t_j= u_1+...+u_j$. Since the population size, N, is unknown, it is convenient to use the notation $P\{D_j=d_j|N=n\}$ and $P\{T_j=t_j|N=n\}$ for the probabilities of D_j and T_j, respectively. The reason for this is the fact that the range of T_j (of N) depends strongly on the unobserved value of N (observed value of T_j).

3. MAIN RESULTS

Given the urn model described in the last section, the following probability statements become straightforward:

(i) Given $m_1 \in \mathbf{N}^*$, $P\{U_1=u_1|N=n\}=1$, for any $n\geq m_1=u_1=t_1$, otherwise is equal to

zero; and (ii) For $j>1$ and $m_j \in \mathbf{N}^*$, $P\{U_j=u_j \mid N=n, U_1=u_1, U_2=u_2, \ldots, U_{j-1}=u_{j-1}\}$

$$= \binom{n-t_{j-1}}{u_j}\binom{t_{j-1}}{m_j-u_j}\left\{\binom{n}{m_j}\right\}^{-1}$$

for any $n \geq \max\{m_1,\ldots,m_j\}$ and $\max\{m_1,\ldots,m_j\} \leq t_j \leq \min\{m_1+\ldots+m_j, n\}$, otherwise is equal to zero.

The only difficulty one may have in understanding the above statements is with the restrictions of n and t_j given in (ii). Note however that to assign m_j ($j \geq 1$) balls to m_j distinct urns one must have $n \geq m_j$ for all $j \geq 1$. On the other hand, since t_j is the number of distinct chosen urns up to the j^{th} stage, it must not be smaller than the number of distinct urns chosen in any stage. Also t_j can neither be greater than the total number of urns, n, nor than the maximum possible number of distinct urns up to the j^{th} stage, $m_1+\ldots+m_j$. Finally, it is not difficult to conclude that the sequence $(T_k)_{k \geq 1}$ is a very interesting Markov Chain (given $\{N=n\}$). In fact, it is a submartingale since, for $j>1$,

$$E\{T_j \mid N=n, T_{j-1}=t\} = \left(1-\frac{t}{n}\right)m_j + t \, .$$

(Sen ,1982; (2.3), introduced a related property for the one-by-one case.)

The following important result is a direct consequence of these probability statements. Recall that $m=m_1+\ldots+m_k$, $u_1=t_1=m_1$, $d_k=(u_1,\ldots,u_k)$, $t_j=u_1+\ldots+u_j$, and $u_j \in \{0,1,\ldots,m_j\}$, for $j=2,\ldots,k$.

3.1 Theorem: For all $k \geq 2$ and $n \in \mathbf{N}^*$ such that $n \geq \max\{m_1,\ldots,m_k\}$,

$$P\{D_j=d_j \mid N=n\} = \frac{(t_k)!\binom{n}{t_k}\prod_{j=2}^{k}\binom{t_{j-1}}{m_{j-1}}}{\prod_{j=1}^{k}\binom{n}{m_j}(u_j)!} I_B(t_k) \, ,$$

where $B=\{x \in \mathbf{N}^*;\ \max\{m_1,\ldots,m_k\} \leq x \leq \min\{m,n\}\}$.

The proof of this result is very simple. To obtain the joint distribution of U_1, U_2,..., and U_k (the distribution of D_k), we need only to consider the product of the conditional probabilities introduced by (i) and (ii) above.

The following lemma is a generalization of a result described by Feller (1968), where the case of $m_1=\ldots=m_k=1$ is considered. In fact it indirectly introduces the

distribution of T_k. Let $P_e\{m_1,...,m_k;n\}$ represent the probability that, at the end of the allocation process, exactly e ($\in \mathbf{N}^*$) urns are empty.

3.2 Lemma: For all $k\geq 1$ and $n\in \mathbf{N}^*$ such that $n\geq \max\{m_1,...,m_k\}$,

$$P_e\{m_1,...,m_k;n\}=\frac{\binom{n}{e}}{\prod_{j=1}^{k}\binom{n}{m_j}}\sum_{i=0}^{n-e}(-1)^i\binom{n-e}{i}\prod_{j=1}^{k}\binom{n-e-i}{m_j}\,I_E(e)\ ,$$

where $E=\{x\in \mathbf{N}^*;\ n-\min\{m,n\}\leq x\leq n-\max\{m_1,...,m_k\}\}$.

Proof: For i=1,...,n, let A_i be the event "the i^{th} urn is empty at the end of the allocation process." Hence, for $1\leq k_1\leq...\leq k_i\leq n$, $P\{A_{k_1}\cap...\cap A_{k_i}\,|N=n\}$

$$=\prod_{j=1}^{k}\binom{n-i}{m_j}\left\{\binom{n}{m_j}\right\}^{-1}\ .$$

On the other hand, $P\{A_1\cup...\cup A_n\,|N=n\}=\sum_{i=1}^{k}(-1)^{i-1}\sum_i P\{A_{k_1}\cap...\cap A_{k_i}\,|N=n\}$, where $\sum_i$ indicates the sum over the set $\{(k_1,...,k_i);1\leq k_1\leq...\leq k_i\leq n\}$ which is composed by $\binom{n}{i}$ points. We can then conclude that $P_0\{m_1,...,m_k;n\}$

$$=1-P\{A_1\cup...\cup A_n\,|N=n\}=\sum_{i=0}^{n}(-1)^i\binom{n}{i}\prod_{j=1}^{k}\binom{n-i}{m_j}\left\{\binom{n}{m_j}\right\}^{-1}I(n\leq m)\ ,$$

where $I(n\leq m)$ is the indicator of $n\leq m$. Replacing n-e for n in the above expression, we notice that $P_0\{m_1,...,m_k;n-e\}\prod_{j=1}^{k}\binom{n-e}{m_j}(m_j)!$ is the number of points favorable to the event "exactly e fixed urns are empty at the end of the allocation process." Recall that the total number of possible allocations of m balls in n-e urns is $\prod_{j=1}^{k}\binom{n-e}{m_j}(m_j)!$. Since, among the n urns, there are $\binom{n}{e}$ ways to choose e urns, we finally have $P_e\{m_1,...,m_k;n\}$

$$=P_0\{m_1,...,m_k;n-e\}\binom{n}{e}\prod_{j=1}^{k}\binom{n-e}{m_j}(m_j)!\ ,$$

which concludes the proof. •

The following result is a direct consequence of the above lemma and is the main result of this paper.

3.3 Theorem: For all $k \geq 1$ and $n \in \mathbf{N}^*$ such that $n \geq \max\{m_1,...,m_k\}$,

$$P\{T_k=t \mid N=n\} = \binom{n}{t}\Big\{\prod_{j=1}^{k}\binom{n}{m_j}\Big\}^{-1}\sum_{i=0}^{t}(-1)^{t-i}\binom{t}{i}\prod_{j=1}^{k}\binom{i}{m_j} I_B(t) .$$

To prove this result we only need to note that if **t** is the number of distinct nonempty urns, then (n-t) is the number of empty urns. Hence, a direct application of Lemma 3.2 produces the desired result. Another consequence, relevant for statistical purposes, is stated next.

3.4 Corollary: For inferences about N, the random variable T_k is a sufficient statistic with respect to D_k. The conditional probability of $\{D_k=d_k\}$ given $\{T_k=t\}$ has the following expression:

$P\{D_k=d_k \mid T_k=t\} = P\{D_k=d_k \mid T_k=t, N=n\}$

$$= \Big\{\prod_{j=1}^{k}(u_j)!\Big\}^{-1}\prod_{j=2}^{k}\binom{t_{j-1}}{m_j-u_j}\Big\{\sum_{i=0}^{t}\frac{(-1)^{t-i}}{i!(t-i)!}\prod_{j=1}^{k}\binom{i}{m_j}\Big\}^{-1} I_{\{t\}}(t_k) .$$

(Recall that the last factor is the indicator of $\{T_k=t\}$.)

That T_k is a sufficient statistic follows from Theorem 3.1 and the well-known Factorization Criterion. Equivalently, sufficiency is also a consequence of the fact that the above conditional probability is the same for all possible values of N. This probability is directly obtained from the expressions introduced in Theorem 3.1 and Theorem 3.3.

4. COMMENTS AND CONCLUSION

The factor

$$K(n;t) = \Big\{(n-t)!\prod_{j=1}^{k}\binom{n}{m_j}\Big\}^{-1} n! ,$$

that appears in the probability expressions of D_k and T_k, is called the **likelihood kernel** since it is the smallest factor of these expressions that depend on the value of n, with the remaining ones independent of n. To obtain maximum likelihood estimates and to perform Bayesian analysis, this kernel is the only sample entity that must be considered. In Leite (1986) these statistical methods are discussed in detail.

Finally, notice that another kind of data could be produced by the urn model described above. For instance, consider the vector $(X_0,X_1,...,X_k)$, where X_i $(0\leq i \leq k)$ is the number of urns with exactly i balls at the end of the allocation process. In terms of population units, X_i is the number of individuals captured exactly i times. Recall that $T_k=X_1+...+X_k$ and $X_0=N-T_k$. With respect to these data, **is T_k still a sufficient statistic?** The answer is again yes. Clearly, after the value **t** of T_k has been recorded, all kinds of nonempty urns must be among these **t**, independently of any possible particular value N may assume. Hence, T_k must be sufficient. To formalize this conclusion we state the following result, the proof of which we shall omit since it would follow the same steps of the ones presented here.

4.1 Theorem: For all $k\geq 2$ and $n\in \mathbf{N}^*$ such that $n\geq \max\{m_1,...,m_k\}$,

$$P\{X_1=x_1,...,X_k=x_k \mid N=n\}$$
$$=K(n;t)\Big\{\prod_{j=1}^{k}(m_j)!\,(x_j)!\Big\}^{-1}h(x_1,...,x_k)I_B(t)\ ,$$

where: (a) the elements of $(x_1,...,x_k)$ take values on $\{0,1,...,k\}$ and satisfy the equations $x_1+2x_2+...+kx_k=m$ and $x_1+...+x_k=t$; and (b) $h(x_1,...,x_k)$ is the number of ways in which m balls can randomly be allocated in **t** urns so that x_1 urns receive one ball, x_2 urns receive 2 balls, and so on up to x_k with k balls.

Here also, by a direct application of the factorization criterion, we conclude that T_k is sufficient. To prove the above result one may need to follow Feller (1968) where the one-by-one case is considered.

We have shown that up to a particular stage, say k, the only relevant information about the unknown parameter of interest, N, is contained in T or equivalently in the likelihood kernel. If, in the place of a fixed stopping step, k, one considers a random stopping rule, the above kernel still would be the minimum sufficient statistic. For example, analogously to the negative binomial rule, suppose that **t** is fixed a priori and k is the number of steps required to obtain **t**. In terms of randomness, k and **t** would change roles; that is, k would be the observation of a random variable and **t** would be the fixed constant. Hence, any desirable good inference about N must rely on a painstaking analysis of the

likelihood kernel, K(n;t). If a random stopping rule is used, instead of CMRR, the sampling scheme is called Capture/Recapture sampling process.

ACKNOWLEDGEMENTS

The authors are grateful to Professors R.E. Barlow and L.R. Pericchi for their comments and suggestions, particularly in the conceptual phase of this paper. The second author would also like to acknowledge support from CNPq, Brasilia, Brazil (contract #20771/85-MA), and the U.S. Army Research Office (contract #DAAG-85-K-0208) during his visit at University of California, Berkeley.

BIBLIOGRAPHY

CRAIG, C.C. 1953. On utilization of marked specimens in estimating the population of flying insects. *Biometrika 4:* 170-6.

GOOD, I.J. 1950. *Probability and the weighing of evidence.* London, Charles Griffin. 119pp.

GOODMAN, L.A. 1953. Sequential sampling tagging for population size problems. *Ann. Math. Statist. 24:* 56-69.

FELLER, W. 1968. *An introduction to probability theory and its applications.* 3 ed. New York, John Wiley. 509pp.

JOHNSON, N.L. & KOTZ, S. 1977. *Urn models and their applications: an approach to modern discrete probability theory.* New York, John Wiley. 402pp.

LEITE, J.G. 1986. *Exact estimates of the size of a finite and closed population* (in Portuguese). Doctoral dissertation. São Paulo, Brazil, Universidade de São Paulo. 93pp.

SAMUEL, E. 1968. Sequential maximum likelihood estimator of the size of a population. *Ann. Math. Statist. 39:* 1057-68.

SEBER, G.A.F. 1986. A review of estimating animal abundance. *Biometrics 42:* 267-92.

SEN, P.K. 1982. A renewal theorem for an urn model. *Ann. Probability 10:* 838-43.

END

10-87

DTIC